Web Personalization Using Web Mining

[1]G.Divya, [2]Shaik Basha

[1]((Electronics and communication Engineering, CMR Engineering College, JNTUH, India)
[2]((Electronics and communication Engineering, CMR Engineering College, JNTUH, India)

Abstract : *This Paper comprises of upcoming field of "Web Mining". Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. In this paper we encompass various Pros and Cons of Web Mining and also represent the similarities and differences between Data Mining and Web Mining. Here we describe some of the recently used Tools used in Web Mining.*
Keywords: *Web Mining, Data Extraction, Web Usage Mining, Web Content Mining, Web Structure Mining*

I.   Introduction

Information has been a important part of humans but with evolution in technology handling of information changed from analog to digital, now we use huge computation to manage our Information and internet has now became a medium for collecting information. Upcoming ERA comprise totally of Internet, expansion of World Wide Web has increased the storing capacity of information online.

T he World Wide Web (WWW) is an information space where documents and other web resources are identified by URLs, interlinked by hypertext links, and can be accessed via the Internet. It consist of different type of Data and storing huge heterogeneous Data online, reverts a serious issue of Extracting that information at the required time.

The term Web Data Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest.

II.   Disadvantages Of Exiting System

The concerns about the personal privacy have been increasing enormously recently especially when the internet is booming with social networks, e-commerce, forums, blogs. Although companies have a lot of personal information about us available online, they do not have sufficient security systems in place to protect that information. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people.
The disadvantages are
 i.   User privacy/Security
 ii.   Amount of data is overwhelming
iii.   Great cost at implementation stage
 iv.   Possible misuse of information
 v.   Possible in accuracy of data

III.   Web Mining

Web mining - is the application of data mining techniques to discover patterns from the World Wide Web. Data Mining is done through various types of data mining software. These can be simple data mining software or highly specific for detailed and extensive tasks that will be sifting through more information to pick out finer bits of information. For example, if a company is looking for information on doctors including their emails, fax, telephone, location, etc., this information can be mined through one of these data mining software programs. This information collection through data mining has allowed companies to make thousands and thousands of dollars in revenues by being able to better use the internet to gain business intelligence that helps companies make vital business decisions.

The data mining also empowers companies to keep a record of fraudulent payments which can all be researched and studied through data mining. This information can help develop more advanced and protective methods that can be undertaken to prevent such events from happening.

The data mining technology is going through a huge evolution and new and better techniques are made available all the time to gather whatever information is required. Web data mining technology is opening avenues on not just gathering data but it is also raising a lot of concerns related to data security.

IV. Web Mining Categories

A. Web Usage Mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.
Web usage mining itself can be classified further depending on the kind of usage data considered:

Web Server Data: The user logs are collected by the Web server. Typical data includes IP address, page reference and access time.

Application Server Data: Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above.

B. Web Content Mining

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. The heterogeneity and the lack of structure that permits much of the ever-expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and search and indexing tools of the Internet.

C. Web Structure Mining

Web structure mining focuses on the hyperlink structure of the Web. The different objects are linked in some way. Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models
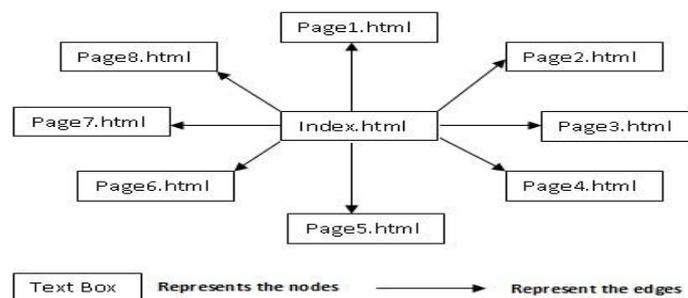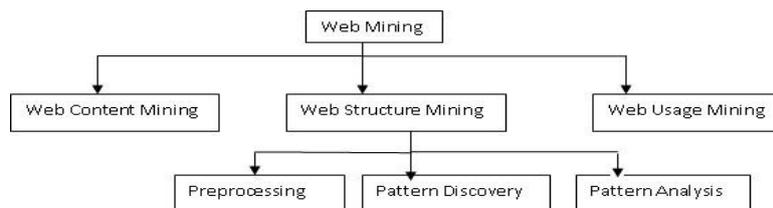
Figure No 2.

Figure No 3.(Components Web Mining)

V. Structure Of Web Mining

Structure of Web Mining represents actually How Web Mining take place. Patterns are evaluated by three techniques of Web Mining i.e Web Content Mining, Web Usage Mining, Web Structure Mining. These techniques evaluate the needed patterns and then these patterns are analyzed to get a user desired output. Desired output is feed into the user understandable GUI [6].
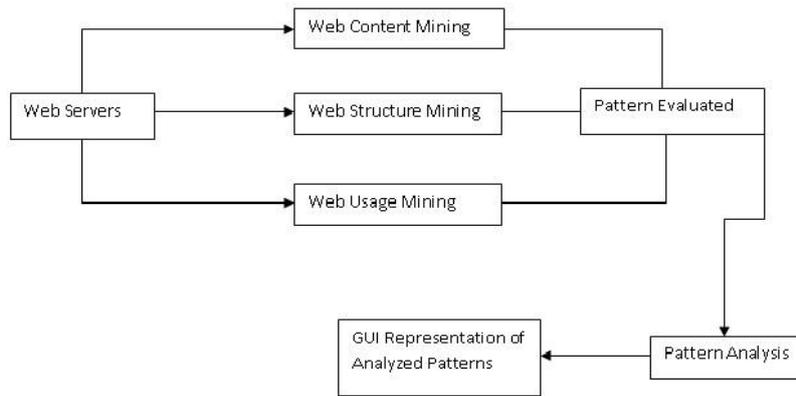
Figure No 4.(Stucture of Web mining)

VI.   Pros And Cons Of Web Mining

Pros
- Web Mining has been a new face of e- commerce due to an ease in personalizing marketing which provides eventually higher volumes of trade.
- It also helps in countering the Cyber Terrorism.
- Web Mining is a powerful tool for Cloud Computing.

Cons
- Web Mining has been a new face of e- commerce due to an ease in personalizing marketing which provides eventually higher volumes of trade.
- It also helps in countering the Cyber Terrorism.
- Web Mining is a powerful tool for Cloud Computing.

VII.   Comparison Between Web Mining And Data Mining

Parameter	Data Mining	Web Mining
Time Stamp	Crawling process take heavy time due to large sizes of databases	Crawling process takes pretty less time due to small sizes of database
Confidentiality	Confidentiality is not rendered because of limited databases in corporate Sector	Confidentiality is render because there are large no. of database online

Table 1 Data Mining v/s Web Mining

VIII.   Tools For Web Mining

Some of useful tools used for Web Mining are [5].
i. **QL2 Software**, the leading provider of data extraction solutions for business and competitive intelligence today   announced the release of WebQL 3.1, the company's sophisticated development platform that simplifies complex data collection and integration tasks and functions. WebQL extracts information from a variety of unstructured data sources, including the Web, and reformats it into structured, more usable formats. The latest release of WebQL has a number of new features, including new URL schemes that provide enhanced flexibility when accessing data sources that are external to WebQL, as well as support for XML data of arbitrary size, a C API for pure C embedding applications and improvements to the network monitor.
ii. **Screen Scraper**, programmatic collection of visual data from a source, instead of parsing data as in Web scraping. Originally, screen scraping referred to the practice of reading text data from a computer displayterminal's screen. This was generally done by reading the terminal's memory through its auxiliary port, or by connecting the terminal output port of one computer system to an input port on another.
iii. **SAS Enterprise Miner**, an integrated suite which provides a user-friendly GUI front-end to the SEMMA (Sample, Explore, Modify, Model, Assess) process.

iv. **SPAD**, provides powerful exploratory analyses and data mining tools, including PCA, clustering, interactive decision trees, discriminate analyses, neural networks, text mining and more, all via user-friendly GUI.

v. **Website Parser**, A web site parser tool is a program that will allow you to gather information from many web sites and web pages throughout the Internet. The tool goes through the targeted sites and is able to grab large amounts of data, through the parameters that you have created. This data can be used in XLS, CSV, XML, and TSV files for later use. Being able to gather huge amounts of information quickly and easily is an invaluable tool for any business owner or retail site.

vi. **Web Extractor Software**, Web extractor software may be one of the smartest software tools to invest in. The cutting edge technology may be used in a variety of settings. It has been effectively utilized by law enforcement, researchers, and several businesses by extracting vital information from specific websites. Data extraction, screen scraping, and web crawling may only be a few of the features available.

vii. **Mozenda-Mozenda**, is a Software as a Service (SaaS) company that enables users of all types to easily and affordably extract and manage web data. With Mozenda, users can set up agents that routinely extract data, store data, and publish data to multiple destinations. Once information is in the Mozenda systems users can format, repurpose, and mashup the data to be used in other online/offline applications or as intelligence.

viii. **WizSoft Software**, Develops software based on mathematical algorithms, mainly for the business sector in the fields of data mining, data auditing, concept-based text search engines, knowledge management, computational linguistics, accounting and inventory management, and operations research.

IX.   Conclusion

This Paper provides a complete and clear cut view of every aspects of Web Mining: Extracting Data Online. We tried to explain Web mining in a simpler way just to make the mass aware of this upcoming technology. This paper clearly depicts the future potential of Web Mining and also provides basic information about Web Mining which is essential for a beginner. The paper also consists of various Diagrams so as to help beginners in understanding the Framework of Web Mining.

References

[1]. A. A. Barfourosh, H.R. Motahary Nezhad, M. L. Anderson, D. Perlis, "*Information Retrieval on the World Wide Web and Active Logic*: A Survey and Problem Definition", *2002.*

[2]. Cooley, R.; Mobasher, B.; Srivastava, J.; "*Web mining: information and pattern discovery on the World Wide Web*". *In Proceedings of Ninth IEEE* International Conference. pp. *558 –567, 3-8 Nov. 1997.*

[3]. G. Shrivastava, K. Sharma, V. Kumar," *Web Mining: Today and Tomorrow*", in the Proceedings of 2011 3rd International Conference on Electronics Computer Technology (ICECT), *pp.399-403, April 2011*

[4]. B. Singh, H.K. Singh, "*Web data Mining Research*", in the Proceedings of 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-10, Dec. 2010

[5]. O. Etzioni, "*The world wide Web: Quagmire or gold mine.* "Communications of the ACM*, Vol. 39*No. 11, pp. 65 -68, Nov.1996.*

[6]. Web mining definition, available: http://en.wikipedia.org/wiki/Web_mining.

[7]. R. Kosala, H. Blockeel "*Web mining research: A survey," ACM SIGKDD Explorations, Vol. 2 No. 1, pp. 1-15, June 2000.*